Distinguishing inherent and contact-influenced intelligibility: An example from the Dejing dialect area of Zhuang

Eric M. Jackson[1]
SIL International, East Asia Group
21st Annual Meeting of the Southeast Asian Linguistics Society, 11-13 May 2011

This talk presents the method used in a recent survey of the Dejing dialect area of Zhuang for distinguishing the influence of previous contact on estimates of dialect intelligibility. It shows the usefulness of comparing test participants' self-reporting of language exposure, results of comprehension testing, and phonetic similarity from word lists.

1 Intelligibility of language varieties

1.1 Why does intelligibility matter?

(1) **Dialectology**: intelligibility is one criterion for classifying speech varieties as related dialects or distinct languages—cf ISO 639-3 criteria (ISO639 2011)

(2) **Language planning and language development**: need to know

- *synchronic sound relationships* for making a multi-dialect alphabetic orthography
- *synchronic intelligibility relationships* for audio & video media sharing

1.2 How can intelligibility of speech varieties be ~~measured~~ estimated?

(3) **Word lists**: similarity by methods like lexicostatistics (Simons 1979, Nahhas & Mann 2006) or string edit distance (Beijering et al 2008, Yang & Castro 2008)

- Relatively easy to do
- Allows comparison with published lists for other related languages

(4) Problems with word lists

- Intelligibility estimate is non-directional
- How does difference quantitatively relate to intelligibility? (Huffman 1976)
  - $<$70% lexical similarity taken as strong indication of mutual unintelligibility
  - Indicates unintelligibility only; can't say what level of similarity guarantees intelligibility

(5) **Comprehension tests**: comprehension scores based on hearing short samples of a speech variety (Casad 1974, Blair 1990, Nahhas 2006)

- Intelligibility estimate is directional
- Results are repeatable, reproducible

---

1 Even though only one person is giving this talk, the research described in this talk is by no means the work of just one person. The fieldwork was carried out by a team of SIL linguists consisting of the author, Lau Shuh Huey, and Emily Jackson, in cooperation with the Guangxi Minorities Language and Scripts Work Commission and its county-level affiliates. For a full account of this work, see Jackson et al (forthcoming).

(6) Problems with comprehension tests

- Hard to control—differences in speaker, content, mean that any two stories are not equally easy to understand
- Comprehension score can reflect multiple sources of intelligibility
  - *inherent intelligibility*: intelligibility of the reference speech variety among speakers of the test speech variety based only on their knowledge of the test variety—marks the lowest bound of intelligibility in a community
  - *acquired intelligibility*: past exposure to the reference speech variety can increase participants' comprehension scores—amount can vary between individuals in a community

(7) For dialectology and language planning & development, inherent intelligibility more useful than inherent + acquired intelligibility together

(8) Inherent intelligibility typically isolated from acquired intelligibility by

- Pre-screening test participants based on degree of past exposure
- Checking final results for low standard deviation (high σ = past exposure is likely)

(9) What about multi-lingual communities where exposure is unavoidable?

2 Intelligibility survey of the Dejing Zhuang dialect area

2.1 Brief description of the survey

(10) Cooperative survey: SIL International and Guangxi Minorities Language and Scripts Working Group

- Jingxi, Napo & Debao Counties, Guangxi Zhuang Autonomous Region; about 1 million people, 97% Zhuang
- Fieldwork carried out in 2008
- Methods, results documented in Jackson et al (forthcoming)

(11) Primary goal: evaluate the suitability of one prominent speech variety (the Yang Zhuang [zyg] of urban Jingxi County) for broad language development, and the scope of other local varieties over which it could be used

(12) Tools used in this dialect survey

- Word lists compared using Levenshtein Distance (LD) by RuG/L04 (Kleiweg 2008)
- Comprehension tests using Recorded Text Test, Re-telling method (Nahhas 2006)
- Participant screening questionnaires (for screening and to gather attitude data)
- Individual and group sociolinguistic questionnaires

(13) In comprehension testing, we always screened for low exposure to the reference variety—but not always possible to screen down to zero exposure

2.2 Our method: compare word lists, comprehension tests, screening

(14) Highly similar word lists *likely* have higher intelligibility (Beijering et al 2008, Castro & Yang 2009, Yang 2010 find high correlation of intelligibility and LD similarity)

(15) Mismatches of predicted intelligibility from word list similarity and measured comprehension point to *potential* influence of acquired intelligibility

(16) Compare screening information for those points, especially responses to the question "What language(s) do you speak?"

- Responses that reflect the reference speech variety and test speech variety as separate varieties spoken by the participant suggest
  - ➢ significant linguistic difference, from the perspective of test participants, and
  - ➢ significant exposure—enough to warrant saying "I speak that variety"

(17) Strongest evidence for influence of acquired intelligibility where all three factors co-occur

- low intelligibility predicted from word lists
- high estimated intelligibility from comprehension testing
- test participants report high exposure, especially referring to the reference variety as separate from the test variety

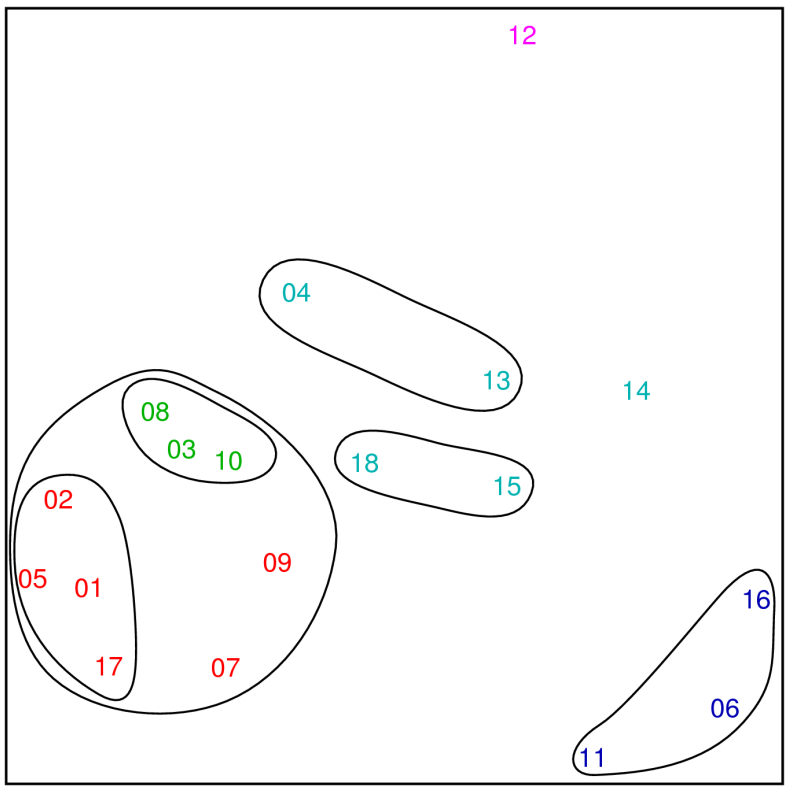
(18) Ideal circumstances for confirming this conclusion are when multiple data locations differ primarily in degree of exposure to the reference variety

- highly similar word lists
- at least one location with very low exposure to the reference speech variety
- significant divergence in comprehension scores indicates likely influence of acquired intelligibility

2.3 Examples from this survey

(19) Word list similarity results: consistent clustering (found by all clustering algorithms)

(20) Word list results: multi-dimensional scale plot



Case 1: Min & Zong Zhuang [zgm] (datapoints 6, 11, 16)

(21) Word list analysis (see (19), (20)): these locations similar to each other, different from Yang reference varieties (01, 02)

(22) Comprehension testing results: moderately high for 6, 11, low for 16 with high σ

Subject Location	Yang reference 01	Yang reference 02
06	84% (σ = 8.6%, n = 8)	94% (σ = 5.4%, n = 8)
11	86% (σ = 11.3%, n = 10)	96% (σ = 3.4%, n = 10)
16	67% (σ = 9.6%, n = 9)	75% (σ = 14.3%, n = 9)

(23) Screening responses

- Datapoint 6: moderate exposure reported; 1 of 8 report speaking local + Yang
- Datapoint 11: high exposure reported; 8 of 10 report speaking local + Yang
- Datapoint 16: low exposure reported; 0 of 9 report speaking local + Yang

(24) Our conclusions

- inherent intelligibility of Yang by speakers of Min & Zong likely low (as reflected by datapoint 16)
- acquired intelligibility in some areas (6, 11) high based on high degree of contact

Case 2: Nong'an Zhuang (datapoint 12)

(25) Word list analysis: this point most different from Yang reference varieties (01, 02)

(26) Comprehension testing results: moderate to high with wide-ranging σ

Subject Location	Yang reference 01	Yang reference 02
12	90% (σ = 11.3%, n = 10)	98% (σ = 2.4%, n = 10)

- Comprehension scores comparable to core Yang cluster (eg, datapoint 5), significantly better than scores at datapoints 4, 13

(27) Screening responses: high exposure reported; 10 of 10 report speaking local + Yang

(28) Our conclusions

- inherent intelligibility of Yang by speakers of Nong'an likely low
- acquired intelligibility high based on high degree of contact

3 Conclusion

(29) Three types of information to collect

- *Phonetic similarity*: (eg, word lists) tends to correlate with (inherent) intelligibility
- *Comprehension results*: estimate of overall intelligibility (inherent and acquired)
- *Screening information*: including amount of exposure to the reference variety and answers to a question like "what language(s) do you speak?"

(30) Complementary evidence doesn't always point clearly to acquired intelligibility

(31) **BUT…** Complementary evidence all pointing in the same direction can increase our confidence that inherent intelligibility really is high or low

References

Beijering, Karin, Charlotte Gooskens & Wilbert Heeringa. 2008. Predicting intelligibility and perceived linguistic distances by means of the Levenshtein algorithm . In Marjo van Koppen and Bert Botma (eds.), *Linguistics in the Netherlands 2008*, 13-24. Amsterdam: John Benjamins Publishing Company.

Blair, Frank. 1990. *Survey on a shoestring*. Dallas: SIL International.

Casad, Eugene. 1974. *Dialect intelligibility testing* (Summer Institute of Linguistics Publications in Linguistics and Related Fields 38). Norman: Summer Institute of Linguistics of the University of Oklahoma.

Huffman, Franklin. 1976. The relevance of lexicostatistics to Mon-Khmer languages. In Jenner, Philip, Laurence Thompson & Stanley Starosta (eds.), *Austroasiatic Studies* (Oceanic Linguistics special publication no. 13), 539-574. Honolulu: University of Hawaii Press.

ISO 639-3 Registration Authority [ISO639]. 2011. Scope of denotation for language identifiers. http://www.sil.org/iso639-3/scope.asp (rechecked 3 May 2011)

Jackson, Eric, Emily Jackson & Shuh Huey Lau. Forthcoming. A sociolinguistic survey of the Dejing Zhuang sub-dialect area (SIL Electronic Survey Reports). Draft available from https://mail.link77.net/~eric_jackson@sil.org/research.html.

Kleiweg, Peter. 2008. RuG/L04 (version of 2008/10/15) [computer program]. Retrieved from http://www.let.rug.nl/~kleiweg/L04/ (accessed October 2008 through October 2010).

Luo, Liming, Ma Chaofa, Margaret Milliken, Nong Guangmin, Wu Jun, Du Zaijing, Lu Zhenyu, Lu Xiaoli, Yang Yijie, Chen Fulong, Zhang Haiying, Mo Jiayu, Liang Jinjie, Huang Dawu, Huang Peixing, Huang Rumeng, Huang Quanxi, Stuart Milliken, Qin Minggui, Qin Yaowu, Xie Lanyan, & Meng Yuanyao (eds.). 2005. *Zhuang-Chinese-English dictionary*. Beijing: Nationalities Press.

Nahhas, Ramzi, compiler. 2006. The steps of Recorded Text Testing: A practical guide. Chiang Mai, Thailand: Payap University, ms. Available online at http://li.payap.ac.th/images/stories/survey/Steps%20of%20RTT.pdf

Nahhas, Ramzi & Noel Mann. 2006. The steps of eliciting and analyzing word lists: A practical guide. Payap University, ms. Available online at http://li.payap.ac.th/images/stories/survey/Steps%20of%20Word%20Lists.pdf.

Simons, Gary. 1979. Language variation and limits to communication. Technical Report No. 3. Ithaca, NY: Department of Modern Languages and Linguistics, Cornell University.

Yang, Cathryn. 2010. Lalo regional varieties: phylogeny, dialectometry, and sociolinguistics. Melbourne: La Trobe University dissertation.

Yang, Cathryn & Andy Castro. 2008. Representing tone in Levenshtein distance. *International Journal of Humanities and Arts Computing: Computing and Language Variation* 2. 205-219.

This handout is available for download at *https://mail.link77.net/~eric_jackson@sil.org/research.html*. (0512a)